

# **STA 291**

# **Summer 2008**

## **Lecture 1**

# WELCOME to STA 291!

- Syllabus:  
<http://www.ms.uky.edu/~kkohrs/sta291.html>
- Class Instructor: Keith Kohrs
- [k.kohrs@uky.edu](mailto:k.kohrs@uky.edu)
- Office Hours: 5 PM – 6 PM  
Mondays and Wednesdays, 818 P.O.T.

# Textbook

- Gerard Keller

## **Statistics for Management and Economics**

7th Edition, Duxbury 2004

***Including ThomsonNow Access***

Can be purchased online or at the campus bookstores.

# Topics

- Statistical Terminology
- Descriptive methods
- Probability and distribution functions
- Estimation (confidence intervals)
- Hypothesis testing
- Inferential methods for two samples
- Simple linear regression and correlation

# Most Important Concepts

- Sampling Distribution (Ch. 9)
- Confidence Interval (Ch. 10)
- P-value of a Statistical Test (Ch. 11)
- Regression (Ch. 17)

# ***Grading***

- Midterm Exam #1 (June 23)..... **23%**
- Midterm Exam #2 (July 14)..... **24%**
- Cumulative Final Exam (July 30)..... **28%**
- Homework Online Assignments..... **15%**
- Quizzes..... **10%**

## **Letter Grades**

**A: 90-100%, B: 80-89%, C: 70-79%,  
D: 60-69%, E: 0-59%**

# Please...



# Why Statistics?

- Research in the sciences is getting more quantitative (look at research journals)
- Computers make even complex statistical methods easier to use
  - danger of using inappropriate methods
  - vital to understand a method before using it
- Job market: Most graduates need to be familiar with basic statistical methodology
- *“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”*

*Herbert George Wells (1866–1946)*

# Why Statistics? (contd)

- Newspaper, advertising, surveys...  
many statements contain statistical arguments

Another Example:  
From Pew Research Center, Posted August 2, 2007

**“A Summer of Discontent with Washington”**

As official Washington winds down for its summer holiday, all three branches of government are coming under fire from the American public.

Just 29% approve of the way President Bush is handling his job, and only slightly more, 33%, approve of the job performance of the Democratic leaders of Congress.

Even the U.S. Supreme Court is not immune from the current round of public disaffection: The court's favorable rating has fallen from 72% in January to 57% currently.

Results for this survey are based on telephone interviews conducted under the direction of Schulman, Ronca & Bucuvalas, Inc. among a nationwide sample of 1,503 adults, 18 years of age or older, from July 25-29, 2007.

For results based on the total sample, one can say with 95% confidence that the error attributable to sampling is plus or minus 3 percentage points.

In addition to sampling error, one should bear in mind that question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of opinion polls.

# What does it take to understand the STA 291 material?

- Don't procrastinate
- Logical thinking
- Perseverance
- ...+ see Syllabus

(attend lectures, obtain material when absent, do homework yourself, etc.)

- ***Make sure you get quickly set up and become familiar with the online homework system***

# What is Statistics?

## Methods for Collecting, Describing, Analyzing, and Drawing Conclusions from Data

These methods are used for...

### **Design**

- Planning research studies
- How best to obtain the required data

### **Description**

- Summarizing data
- Exploring patterns in the data
- Extract/condense information
- Graphical pictures of the data

### **Inference**

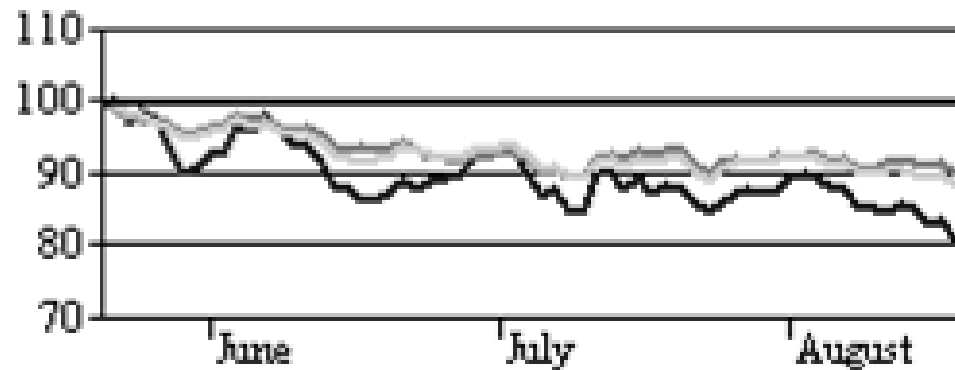
- Make predictions based on the data
- “Infer” from sample to population
- Generalize

# *Descriptive Statistics, e.g.*

Frequency Distribution

Highest Degree	Number of Employees
Grade School	15
High School	200
Bachelor's	185
Master's	55
Doctorate	70
Other	25
Total	550

Time Plot



# Basic Terminology I

- **Population**
  - total set of all subjects of interest
  - the entire group of people, animal or things about which we want information
- **Elementary Unit**
  - any individual member of the population
- **Sample**
  - subset of the population from which the study actually collects information
  - used to draw conclusions about the whole population

# Basic Terminology II

- **Variable**

- a characteristic of a unit that can vary among subjects in the population/sample
- Examples: gender, nationality, age, income, hair color, height, disease status, company rating, grade in STA 291, state of residence

- **Sampling Frame**

- listing of all the units in the population

- **Parameter**

- numerical characteristic of the **p**opulation
- calculated using the whole population

- **Statistic**

- numerical characteristic of the **s**ample
- calculated using the sample

Examples

# Data Collection and Sampling

***Why not measure all of the the units in the population? Why not take a census?***

## **Problems:**

- *Accuracy:* May not be able to list them all.
- *Time:* Speed of Response
- *Expense:* Cost
- *Infinite Population*
- *Destructive Sampling or Testing*

# Flavors of Statistics

- **Descriptive Statistics**
  - Summarizing the information in a collection of data
- **Inferential Statistics**
  - Using information from a sample to make conclusions/predictions about the population

# Example 1

- University Health Services at UK conducts a survey about alcohol abuse among students.
- 200 of the 30,000 students are sampled and asked to complete a questionnaire.
- One question is “have you regretted something you did while drinking”?
- What is the population? Sample?
- For the 30,000 students, of interest is the percentage who would respond “yes”.

This value is computed for the students sampled.

Is this a parameter or a statistic?

# Example 2

- The Current Population Survey of about 60,000 households in the United States in 2002 distinguishes three types of families: Married-couple (MC), Female householder and no husband (FH), Male householder and no wife (MH).
- It indicated that 5.3% of “MC”, 26.5% of “FH”, and 12.1% of “MH” families have annual income below the poverty level.
- Are these numbers statistics or parameters?
- The report says that the percentage of all “FH” families in the USA with income below the poverty level is at least 25.5% but no greater than 27.5%.
- Is this an example of descriptive or inferential statistics?

# Modified Example

- A census of all households in Lexington indicated that 6.2% of married couple households in Lexington have annual income below the poverty level.
- Is this number a statistic or a parameter?

# Univariate vs Multivariate

- Univariate data set
  - Consists of observations on a single attribute
- Multivariate data
  - Consists of observations on several attributes
- Special case: Bivariate data
  - Two attributes collected per observation

# Scales of Measurement

- Qualitative and Quantitative
  - Nominal and Ordinal
  - Discrete and Continuous
- 
- **Recall:**
    - A **Variable** is a characteristic of a unit that can vary among subjects in the population/sample

# Qualitative Variables (=Categorical Variables) Nominal or Ordinal

- **Nominal**: gender, nationality, hair color, state of residence
- Nominal variables have a **scale of unordered categories**
- It does not make sense to say, for example, that green hair is greater/higher/better than orange hair

# Qualitative (Categorical) Variables

## Nominal or Ordinal

- **Ordinal:** Disease status, company rating, grade in STA 291
- Ordinal variables have a scale of ordered categories. They are often treated in a quantitative manner (A=4.0, B=3.0,...)
- One unit can have more of a certain property than does another unit

# Quantitative Variables

- **Quantitative:** age, income, height
- Quantitative variables are measured numerically, that is, for each subject, a number is observed
- The scale for quantitative variables is called **interval scale**

# Example 1

- Vigild (1988) “Oral hygiene and periodontal conditions among 201 institutionalized elderly”, Gerodontology, 4:140-145
- Variables measured
  - Nominal: Requires Assistance from Staff?  
Yes / No
  - Ordinal: Plaque Score  
No Visible Plaque - Small Amounts of Plaque -Moderate Amounts of Plaque - Abundant Plaque
  - Interval: Number of Teeth

## Example 2

- The following data are collected on newborns as part of a birth registry database
- Ethnic background: African-American, Hispanic, Native American, Caucasian, Other
- Infant's Condition: Excellent, Good, Fair, Poor
- Birthweight: in grams
- Number of prenatal visits
  
- What are the appropriate scales?

# Why is it important to distinguish between different types of data?

- Some statistical methods only work for quantitative variables, others are designed for qualitative variables.

Nominal	-	Ordinal	-	Interval
Qualitative (Categorical)				Quantitative
Lowest level				Highest Level
				- most information
				- best statistical methods

You **can not** use statistical methods for quantitative data to analyze qualitative data.

You **can** treat variables in a less quantitative manner.

- Example.

- Height: Quantitative variable, interval scale,  
*measured in cm (or ft/in)*

- Can be treated as ordinal  
*short, average, tall*

- Can even be treated as nominal  
*180cm-200cm, all others*

- Try to measure variables at the highest possible level
- Higher-level variables can be analyzed with a greater variety of methods

Caution: Sometimes, ordinal variables are treated as quantitative

# Discrete and Continuous

- A variable is discrete if it can take on a finite number of values
- Examples: gender, nationality, hair color, disease status, company rating, grade in STA 291, state of residence
- Qualitative (categorical) variables are discrete

# Discrete and Continuous

- Continuous variables can take an *infinite continuum* of possible real number values
- Example: time spent on STA 291 homework
  - can be 63 min. or 85 min.  
or 27.358 min. or 27.35769 min. or ...
  - can be **subdivided**
  - therefore **continuous**

# Discrete or Continuous

- Another example: number of children
- can be 0, 1, 2, 3, ...
- can not be 1.5 or 2.768
- can **not** be **subdivided**
- therefore not continuous but **discrete**

# Discrete or Continuous

- Quantitative variables can be discrete or continuous
- How about age, income, height?
- **It depends** on the scale
- Age is potentially continuous, but usually measured in years (discrete)

- 5.1 Methods of Collecting Data
- 5.2 Sampling
- 5.3 Sampling Plans
- 5.4 Sampling and Nonsampling Errors

# Simple Random Sampling

- Each possible sample has the same probability of being selected.
- The sample size is usually denoted by  $n$ .

# Example: Simple Random Sampling

- Population of 4 students: Adam, Bob, Christina, Dana
- Select a simple random sample (SRS) of size  $n=2$  to ask them about their smoking habits
- 6 possible samples of size  $n=2$ :
  - (1) A+B, (2) A+C, (3) A+D
  - (4) B+C, (5) B+D, (6) C+D

# How to choose a SRS?

- Each of the six possible samples has to have the same probability of being selected
- For example, roll a die (or use a computer-generated random number) and choose the respective sample
- Online Sampling Applet

# How not to choose a SRS?

- Ask Adam and Dana because they are in your office anyway
  - “convenience sample”
- Ask who wants to take part in the survey and take the first two who volunteer
  - “volunteer sampling”

# Problems with Volunteer Samples

- The sample will poorly represent the population
- Misleading conclusions
- BIAS
- Examples: Mall interview, Street corner interview

# Methods of Collecting Data I

## Observational Study

- An observational study observes individuals and measures variables of interest but does not attempt to influence the responses.
- The purpose of an observational study is to describe/compare groups or situations.
- Example: Select a sample of men and women and ask whether he/she has taken aspirin regularly over the past 2 years, and whether he/she had suffered a heart attack over the same period

# Methods of Collecting Data II Experiment

- An experiment deliberately imposes some treatment on individuals in order to observe their responses.
- The purpose of an experiment is to study whether the treatment causes a change in the response.
- Example: Randomly select men and women, divide the sample into two groups. One group would take aspirin daily, the other would not. After 2 years, determine for each group the proportion of people who had suffered a heart attack.

# Methods of Collecting Data III

## Observational Study/Experiment

- **Observational Studies** are passive data collection
- We observe, record, or measure, but don't interfere
- **Experiments** are active data production
- Experiments actively intervene by imposing some treatment in order to see what happens
- *Experiments are preferable if they are possible*
- Examples

# Sampling: Famous Example

- 1936 presidential election
- Alfred Landon vs. Franklin Roosevelt
- Literary Digest sent over 10 million questionnaires in the mail to predict the election outcome
- More than 2 million questionnaires returned
- Literary Digest predicted a landslide victory by Alfred Landon

- George Gallup used a much smaller **random** sample and predicted a clear victory by Franklin Roosevelt
- Roosevelt won with 62% of the vote
- Why was the Literary Digest prediction so far off?

# Other Examples

- TV, radio call-in polls
- “should the UN headquarters continue to be located in the US?”
- ABC poll with 186,000 callers: 67% no
- Scientific random sample with 500 respondents: 28% no
- The smaller **random** sample is much more trustworthy because it has less bias

- Cool inferential statistical methods can be applied to state that “the true percentage of all Americans who want the UN headquarters out of the US is between 24% and 32%”
- These methods **can not** be applied to a volunteer sample.

# How to Choose a Simple Random Sample (SRS)

- Each possible sample has the same probability of being selected.
- The sample size is denoted by  $n$ .
- Enumerate all possible samples, and then randomly choose one of them
- Or, let the computer choose a random sample, for example using this tool:

[Online Sampling Applet](#)

# How not to choose a SRS?

- “convenience samples” or “volunteer samples” like Mall interview or call-in polls
- The sample will poorly represent the population
- Misleading conclusions
- BIAS

# Why are call-in polls usually biased?

- People are much more likely to call in if they feel strongly about an issue  
(Israel-Palestine, Iraq, water company, mountaintop removal, equal rights for homosexuals, pedestrian safety, name of the UK mascot)

# Don't trust bad samples

- Whenever you see results from a poll, check whether they come from a random sample
- Preferably, it should be stated
  - Who sponsored and conducted the poll?
  - How were the questions worded?
  - How was the sample selected?
  - How large was it?
- ***If not, the results may not be trustworthy***

# Question Wording

- Kalton et al. (1978), England
- Two groups get questions with slightly different wording

# Question Wording

- Group 1 is asked: "Are you in favor of giving special priority to buses in the rush hour *or not?*"
- Group 2 is asked: "Are you in favor of giving special priority to buses in the rush hour *or should cars have just as much priority as buses?*"

# Question Wording

- Result: Proportion of people saying that priority should be given to buses.

	Without reference to cars	With reference to cars	Difference
All respondents	<b>0.69</b> (n=1076)	<b>0.55</b> (n=1081)	<b>0.14</b>
Women	<b>0.65</b> (n=585)	<b>0.49</b> (n=590)	<b>0.16</b>
Men	<b>0.74</b> (n=491)	<b>0.66</b> (n=488)	<b>0.08</b>
Non Car-owners	<b>0.73</b> (n=565)	<b>0.55</b> (n=554)	<b>0.18</b>
Car owners	<b>0.66</b> (n=509)	<b>0.54</b> (n=522)	<b>0.12</b>

# Question Order

- Two questions asked in different order during the cold war
- (1) "Do you think the U.S. should let Russian newspaper reporters come here and send back whatever they want?" 36% answered "Yes"
- (2) "Do you think Russia should let American newspaper reporters come in and send back whatever they want?"
- When question (2) was asked first, 73% answered "Yes" to question (1)

# Stratified Sampling

- Suppose the population can be divided into separate, non-overlapping groups ("***strata***") according to some criterion.
- Select a simple random sample independently from each group.

# Why could stratification be useful?

- We may want to draw inference about population parameters for each subgroup
- Sometimes, (“proportional stratified sample”) estimators from stratified random samples are more precise than those from simple random samples

# Proportional Stratification

- The proportions of the different strata are the same in the sample as in the population
- Mathematically:

Population size  $N$ , subpopulation sizes  $N_i$

Sample size  $n$ , subsample sizes  $n_i$

$$\frac{n_i}{n} = \frac{N_i}{N}$$

# Proportional Stratification

- Example:
  - Total population of the US: 303 Million
  - Population of Kentucky: 4 Million (1.3%)
  - Suppose you take a sample of size  $n=303$  of people living in the US.
  - If stratification is proportional, then 4 people in the sample need to be from Kentucky
  - Suppose you take a sample of size  $n=1000$ . If you want it to be proportional, then 13 people (1.3%) need to be from Kentucky.

# Cluster Sampling

- The population can be divided into a set of non-overlapping subgroups (the clusters)
- The clusters are then selected at random, and all individuals in the selected clusters are included in the sample
- Example

# Systematic Sampling

- An initial name is selected at random
- every  $K$ th name is thereafter selected
- $K$  is computed by dividing membership list length by the desired sample size
- Not a simple random sample (why?), but often almost as good as one
- Example

# Summary of Important Sampling Plans

- **Simple Random Sampling (SRS)**
  - Each possible sample has the same probability of being selected.
- **Stratified Random Sampling**
  - Non-overlapping subgroups (strata)
  - SRSs are drawn from each strata
- **Cluster Sampling**
  - Non-overlapping subgroups (clusters)
  - Clusters selected at random
  - All individuals in the selected clusters are included in the sample
- **Systematic Sampling**
  - Useful when the population consists as a list
  - A value  $K$  is specified. Then one of the first  $K$  individuals is selected at random, after which every  $K$ th observation is included in the sample

# Types of Bias

- **Selection Bias**
  - Selection of the sample systematically excludes some part of the population of interest
- **Measurement/Response Bias**
  - Method of observation tends to produce values that systematically differ from the true value
- **Nonresponse Bias**
  - Occurs when responses are not actually obtained from all individuals selected for inclusion in the sample

# Biased or Unbiased Sample?

- Researchers state, "This study was conducted at a large, predominantly White southwestern university. On this campus, American Indians were the smallest racial and ethnic minority student group, consisting of only 2.3% of the student population. Recruited through education and liberal arts classes, students who volunteered to participate in this study completed the research packet and returned it during the next class period. A total of 83 American Indian undergraduates returned completed survey packets."

Gloria, Kurpius (2001), *Cultural Diversity and Ethnic Minority Psychology*, 7, 88-102

# Next Definition: Sampling Error

- Assume you take a random sample of 100 UK students and ask them about their political affiliation (Democrat, Republican, Independent)
- Now take another random sample of 100 UK students
- Will you get the same percentages?

- No, because of sampling variability.
- Also, the result will not be exactly the same as the population percentage, unless you take a “sample” consisting of the whole population of 30,000 students (this would be called a “census”)

or if you are very lucky

# Sampling Error

- **Sampling Error** is the error that occurs when a statistic based on a sample estimates or predicts the value of a population parameter.
- In random samples, the sampling error can usually be quantified.
- In nonrandom samples, there is also sampling variability, but its extent is not predictable.

# Nonsampling Error

- Everything that could also happen in a census, that is, when you ask the whole population
- Examples: Bias due to question wording, question order, nonresponse (people refuse to answer), wrong answers (especially to delicate questions)